PROTEOMIC ANALYSIS OF *E. COLI* USING 2D HPLC AND MALDI-TOF MASS SPECTROMETRY

A Senior Thesis

By

CHRISTOPHER S. CAMPBELL

Submitted to the Office of Honors Programs

& Academic Scholarships

Texas A&M University

In partial fulfillment of the requirements of the

UNIVERSITY UNDERGRADUATE

RESEARCH FELLOWS

April 2002

Group:

Life Sciences 1

PROTEOMIC ANALYSIS OF *E. COLI* USING 2D HPLC AND MALDI-TOF MASS SPECTROMETRY

A Senior Thesis

By

CHRISTOPHER S. CAMPBELL

Submitted to the Office of Honors Programs

& Academic Scholarships

Texas A&M University

In partial fulfillment of the requirements of the

UNIVERSITY UNDERGRADUATE

RESEARCH FELLOWS

Approved as to style and content by:

James C. Hu

Edward A. Funkhouser

April 2002

Group:

Life Sciences 1

ABSTRACT

Proteomic Analysis of *E. coli* Using 2D HPLC and MALDI-TOF Mass Spectrometry.

Christopher S. Campbell

Department of Biochemisty/Biophysics

Texac A&M University

Fellows Advisor: Dr. James C. Hu

Department of Biochemisty/Biophysics

In this post-genomic era, researchers are striving to find new ways to use the enormous amounts of data that have been collected. One obvious way is with proteomics, the large-scale identification of expressed proteins. We have developed a novel method for identifying proteins using two dimensions of non-denaturing high performance liquid chromatography (HPLC) and matrix assisted laser desorption ionization time of flight (MALDI-TOF) mass spectrometry. The first dimension of separation uses an anion exchange column and each of those fractions is run through the second dimension, a hydrophobic interaction column. The proteins were then dialyzed, denatured, and digested with trypsin before being subjected to mass spectrometry. Identifications were made based on the peptide masses. Using this method we have made 2012 protein identifications, 310 of which are unique. These numbers are comparable to other forms of proteomics such as 2-D gels.

This thesis is dedicated to Jimmy.

I'll always remember you Jimmy.

ACKNOWLEDGEMENTS

I would like to thank Dr. Hu and Matthew Champion for their guidance. In addition, I would like to thank Matt for help with the figures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Proteomics is one if the most rapidly developing fields in the biological sciences. As such, a quick and efficient method for performing proteomic analyses is essential. The traditional method for proteomics involves the use of 2-D gels to visualize the proteins. 2-D gels first separate proteins based on their pI using isoelectric focusing. The second mode of separation is sodium dodecyl sulfate polyacrylamide gel electrophoresis. The spots are then excised and identified with mass spectrometry. However, there have been numerous complaints brought up against 2-D gels. Low abundance proteins are difficult to identify on 2-D gels. Many proteins have similar pIs, making separations difficult.[1] 2-D gels also have an apparent bias towards proteins in the lower pI ranges. An alternate method that has been gaining popularity is the use of various forms of liquid chromatography for separating the proteins.[2,3] Often affinity or reverse phase chromatography is used. These methods also use mass spectrometry to make identifications. There are problems with these methods as well. They are very costly and labor intensive. They also have an excessive false positive rate upwards of 30 percent.[4] My thesis research involved developing an alternate method. Our method uses two forms of non-denaturing liquid chromatography in series; anion exchange and hydrophobic interaction. This separates the proteins into 380 fractions each containing 0-7 identifiable proteins. The fractions are each dialyzed, denatured and digested with trypsin before being subjected to analysis by MALDI-TOF mass spectrometry. Identifications are made using Protein Prospector

MS-Fit software. Our method can be done with very low cost and only one or two people. We did the complete experiment four times.

Proteomics involves more than just identifications. Ideally, we would like to know what the proteins are interacting with. Since the mode of separation is non-denaturing, protein activity and complexes are preserved. We have proven β-galactosidase activity is maintained through both chromatography steps. Many, if not most proteins are believed to exist in complexes. Experiments such as those done by Ho et al.[5] and Eisenberg et al.[6] have been able to provide some evidence for the existence of specific complexes. However, multiple types of experiments are needed to get the whole picture. We hope to be able to identify protein complex candidates based on co-fractionation. The entire proteomic analysis has been done with 2 different pHs at which the anion exchange is run. Changing the pH changes which fractions the proteins elute into. By observing which proteins continue to co-fractionate after the shift, we have been be able to accumulate more circumstantial evidence for the existence of complexes.

MATERIALS AND METHODS

One liter of *E. coli* (MG1655) cells were grown in minimal glucose (M9) media. Cells were pelleted at 4000g for 20 minutes and resuspended in 200ml of 20mM Tris-HCl, 20mM NaCl, 1mM EDTA, pH 8.75. They were then centrifuged again and resuspended in 6ml of the same buffer. The cells were lysed via French press and half of the lysate was loaded onto a 1ml Waters column packed with SOURCE 15 Q resin. A gradient from 20mM to 1M NaCl was used. The pH set for the run was either 7.5 or 8.75. Five ml fractions were collected and each of them was loaded onto a 1ml Waters column packed with SOURCE 15 Phe resin. The gradient used went from 1.5M to 0M ammonium sulfate. Each 500µl fraction was collected directly into a Slide-A-Lyzer MINI Dialysis unit (Pierce 3,500 MWCO). The samples were dialyzed for 24 hours in 25mM ammonium bicarbonate. They were then denatured with heat (95°F for 20') and digested with 1µg of modified trypsin (Promega) each for 5 hours. The MALDI was done in a similar fashion to that previously described by Park et al..[7] Identifications from the peptide mass data were made with Protein Prospector MS-Fit software (prospector.ucsf.edu). Factors looked at for making the identifications include MOWSE score,[8] sequence coverage, number of peptides matched, and trends in peptide error. All of the identified proteins were checked to make sure that they are present in the genomic DNA sequence of *E. coli* strain MG1655.[9]

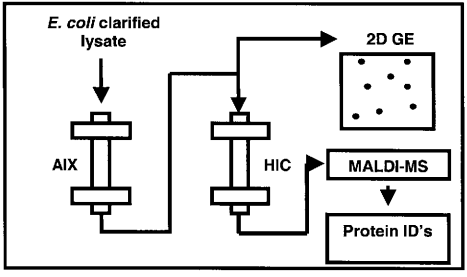


Figure 1. Flow chart describing the path of the proteins for identification by MALDI and for 2D gel analysis.

RESULTS AND CONCLUSIONS

A total of 2012 identifications have been made which include 310 different proteins identified. This is slightly more than the 271 different proteins identified by the Swiss 2-D project.[10,11] Figure 2 shows the overlap between the two projects.

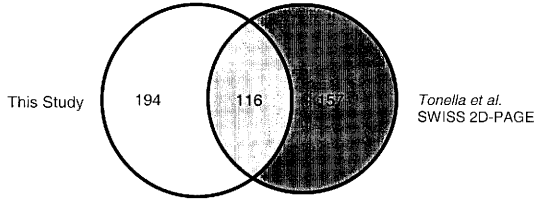
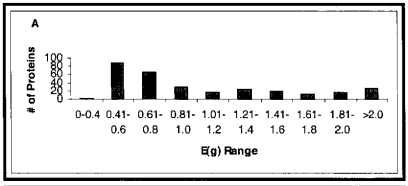


Figure 2. Venn Diagram showing the number of proteins found in only our study, only the SWISS 2D-PAGE project, and those found in both.

One thing that we wanted to determine was whether or not we had any biases in our identifications towards things with high abundance or high/low pI or molecular weight. To measure the abundance of the proteins we identified, we used E(g) numbers. E(g) numbers predict protein abundance based on the codon usage of their genes.[12] The higher the number, the more abundant the protein is predicted to be. On average, our E(g) numbers were significantly higher than those of the entire predicted proteome of *E. coli*, indicating that we do have a bias towards proteins of higher abundance. Figure 3 shows a comparison between the percentage of proteins in different E(g) ranges for our data and the entire proteome.

Our Data



Genome

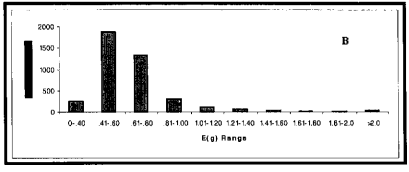


Figure 3. Distribution of proteins for a range of E(g) numbers for A) the proteins that we have identified and B) the predicted proteome.

However, our numbers were still lower than those of other proteome projects. Figure 4 compares our E(g) numbers to those of other projects.

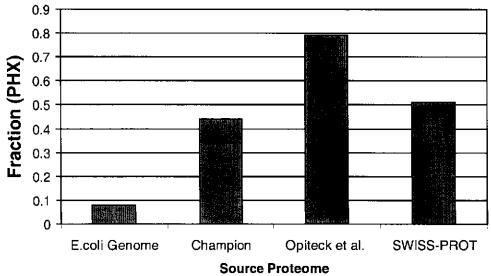


Figure 4. Comparison of average E(g) number for a variety of protein sets, including the *E. coli* genome and three proteomics projects.

To check for pI and molecular weight biases, we simply compared the theoretical pIs and molecular weights for the proteins we identified and the entire genome. These were calculated using the prediction tools found on the SWISS-PROT website (www.expasy.org). There did not appear to be any significant difference between our pI and molecular weight distribution and that of the proteome. Figure 5 shows graphs of pI vs. molecular weight for our data as well as the proteome.

To identify candidates for protein complexes, we found all pairs of proteins that were found in the same fraction for both pH 7.5 and 8.75. Using this method, 125 candidate interactions were identified (Table 1). In addition, some known complexes were found to co-fractionate. For example, phenylalanine tRNA synthetase α and β subunits were found together. RNA polymerase subunits α, β, and β' were also seen in the same fractions. We believe that this data in conjunction with other experiments

similar to those done by Ho et al. and Eisenberg et al. could result in fairly confident identification of novel complexes.

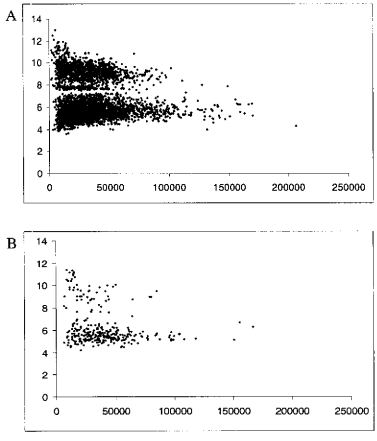


Figure 5. Graphs of pI versus molecular weight for A) the predicted proteome and B) the proteins that we identified.

ACEA	PNP	DAPD	PURT	GREA	GUAA	PROA	GLNS	SUCC	CYSK
ACKA	FABI	DAPD	SSPA	GREA	PPA	PROS	FABI	TALB	PYRH
ACKA	TSF	DNAK	LYSS	GROS	TIG	PROS	PURH	THRC	ASPC
ADK	GAPA	DNAK	TYPA	GUAA	DAPA	PROS	TSF	TIG	ASNS
AHPC	GLNS	DUT	GND	GUAA	GREA	PROS	TUFA	TIG	GROS
AHPC	TRPC	ENO	GND	GUAA	PPA	PURA	GLTA	TIG	GUAB
ALAS	YADF	ENO	SERC	GUAA	YCHF	PURA	KBL	TIG	PHES
ARGD	FUSA	FABI	ACKA	GUAB	TIG	PURA	TKTA	TIG	PHET
ARGG	ISCS	FABI	PROS	HISC	YADF	PURF	ARGG	TIG	RFBB
ARGG	PURF	FABI	PURH	ILES	ASPC	PURF	ISCS	TIG	RPLJ
ARGH	CLPP	FABI	TSF	INFB	LYSS	PURF	PNP	TIG	RPSA
ARGH	FUSA	FABI	YADF	ISCS	ARGG	PURF	TYPA	TKTA	GLTA
ARGI	GCVT	FDX	LPDA	ISCS	CLPP	PURH	FABI	TKTA	PURA
AROA	DAPD	FUSA	ARGD	ISCS	PNP	PURH	PROS	TKTA	TSF
AROK	CYSK	FUSA	ARGH	ISCS	PURF	PURH	TSF	TKTA	TUFA
AROK	PGI	FUSA	ASNS	ISCS	SLYD	PURH	TUFA	TPIA	GLYA
ASNS	DAPA	FUSA	RPSA	KBL	ASPS	PURH	YADF	TRPC	AHPC
ASNS	FUSA	FUSA	SPEE	KBL	GND	PURN	SSPA	TRPC	GLNS
ASNS	GLTA	FUSA	VALS	KBL	PURA	PURT	DAPD	TSF	ACKA
ASNS	KDGK	GAPA	ADK	KDGK	ASNS	PYKF	CYSK	TSF	FABI
ASNS	RFBB	GAPA	GLYA	KDGK	DAPA	PYKF	GCVT	TSF	GLTA
ASNS	RPLJ	GAPA	GPMA	LPDA	FDX	PYKF	NDK	TSF	PPIB
ASNS	RPSA	GCVT	ARGI	LYSS	DNAK	PYRH	TALB	TSF	PROS
ASNS	SERS	GCVT	ASPS	LYSS	INFB	RFBB	ASNS	TSF	PURH
ASNS	TIG	GCVT	CYSK	NDK	DAPD	RFBB	RPSA	TSF	RPLI
ASNS	TUFA	GCVT	NDK	NDK	GCVT	RFBB	TIG	TSF	TKTA
ASNS	VALS	GCVT	PYKF	NDK	PYKF	RPLI	TSF	TSF	TUFA
ASPC	DAPD	GLNS	AHPC	NUSA	PNP	RPLJ	ASNS	TUFA	ASNS
ASPC	ILES	GLNS	PROA	NUSA	SLYD	RPLJ	TIG	TUFA	GLTA
ASPC	THRC	GLNS	TRPC	NUSA	SPEB	RPSA	ASNS	TUFA	PROS
ASPS	GCVT	GLTA	ASNS	NUSA	YICC	RPSA	FUSA	TUFA	PURH
ASPS	GND	GLTA	PURA	PGI	AROK	RPSA	RFBB	TUFA	TKTA
ASPS	KBL	GLTA	TKTA	PGI	CYSK	RPSA	SERS	TUFA	TSF
BGLA	YFBU	GLTA	TSF	PHES	PHET	RPSA	TIG	TYPA	DNAK
CLPP	ARGH	GLTA	TUFA	PHES	TIG	RPSA	VALS	TYPA	PNP
CLPP	ISCS	GLTX	GND	PHET	PHES	RSUA	VALS	TYPA	PURF
CYSK	AROK	GLTX	PPA	PHET	TIG	SERC	ENO	VALS	ASNS
CYSK	DAPD	GLYA	GAPA	PNP	ACEA	SERC	GLYA	VALS	FUSA
CYSK	GCVT	GLYA	SERC	PNP	ISCS	SERS	ASNS	VALS	RPSA
CYSK	PGI	GLYA	TPIA	PNP	NUSA	SERS	RPSA	VALS	RSUA
CYSK	PYKF	GLYA	YIFE	PNP	PURF	SLYD	ISCS	YADF	ALAS
CYSK	SUCC	GND	ASPS	PNP	SLYD	SLYD	NUSA	YADF	FABI
DAPA	ASNS	GND	DUT	PNP	TYPA	SLYD	PNP	YADF	HISC
DAPA	GUAA	GND	ENO	PNP	YICC	SLYD	SPEB	YADF	PURH
DAPA	KDGK	GND	GLTX	PPA	DAPA	SLYD	YICC	YCHF	GUAA
DAPA	PPA	GND	GOR	PPA	GLTX	SPEB	NUSA	YFBU	BGLA
DAPD	AROA	GND	KBL	PPA	GND	SPEB	SLYD	YICC	NUSA
DAPD	ASPC	GND	PPA	PPA	GREA	SPEE	FUSA	YICC	PNP
DAPD	CYSK	GOR	GND	PPA	GUAA	SSPA	DAPD	YICC	SLYD
DAPD	NDK	GPMA	GAPA	PPIB	TSF	SSPA	PURN	YIFE	GLYA

Table 1. Proteins that cofractionate at both pH7.5 and pH8.75. The 125 pairs are shown as 250 entries;each pair is listed with each partner first to aid finding proteins of interest.

To find out what kinds of proteins we identified, we looked at their functional classifications. Table 2 compares the percentages of functional classifications for our data as well as the entire genome. The lack of membrane proteins was expected because the membranes are pelleted after the cells are lysed. We identified a higher proportion of protein biosynthesis and nucleotide metabolism proteins most likely because of their high abundance.

Category:	% of Total	% of Genome
Protein Biosynthesis/ Chaperonin	19%	4.5%
Glycolysis TCA Carbon Utilization	15%	13.0
NT Metabolism	11%	1.4%
AA Synthesis	10%	3.0%
Enzymatic Activities	10%	11.9
Biosynthetic Genes FA, DAP, LPS Cofactors	10%	8.7%
Transcription	4%	1.3%
Replication	1%	2.7%
Membrane, Xport	0%	10.3
Hypothetical/ Unknown/ Putative	19%	43.2
Total	100	100

Table 2. Percentage of proteins in each functional classification for the proteins identified in this study and the *E. coli* genome. Classification come from Opiteck et al.[13]

To help confirm our identifications, 2D PAGE was performed on all of the first dimension fractions run at pH 7.5. Our identifications for each fraction were compared to the spots found on the gels. Both of the predicted molecular weight and pI as well as known migration patterns for previously identified proteins were looked at. The proteins identified by us had a much higher chance of correlating with a spot on these gels than proteins selected at random. Spots were assigned to 109 of the 219

proteins we identified at pH 7.5. Forty-one of these have not been identified by the SWISS-2D project.

At least one false positive was confirmed. UDP-glucose dehydrogenase from the K5 strain of *E. coli* was identified. To make sure that there was not something wrong with either our strain of E. coli or the MG1655 UDP-glucose dehydrogenase sequence, we sequenced the gene. The results showed that our strain does indeed have the MG1655 version of UDP-glucose dehydrogenase, and not that of K5. All other identified proteins came from MG1655. However, some Ids may still be false positives.

Future work on this project will likely involve proving that the system can be used to identify differences in protein expression under altered conditions. Repeating the experiment with cells that have been infected with lambda phage and looking for differences from the previous experiments is one possible way of doing this. Another would be to look at protein expression in outgrowth after starvation.

REFERENCES

1. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **97**: 9390-5.

2. Washburn, M.P., Wolters, D., and Yates, J.R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**: 242-7.

3. Wolters, D.A., Washburn, M.P., and Yates, J.R., 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* **73**: 5683-90.

4. Eriksson, J., Chait, B., Fenyo, D. (200) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem* **72**: 999-1005.

5. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**: 180-3.

6. Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. (2000) Protein function in the post-genomic era. *Nature* **405**: 823-826.

7. Park, Z.Y., and Russell, D.H. (2001) Identification of individual proteins in complex protein mixtures by high-resolution, high-mass-accuracy MALDI TOF-mass spectrometry analysis of in-solution thermal denaturation/enzymatic digestion. *Anal Chem* **73**: 2558-64.

8. Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* **3**: 327-332.

9. Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453-1462.

10. Tonella, L., Hoogland, C., Binz, P.A., Appel, R.D., Hochstrasser, D.F., and Sanchez, J.C. (2001) New perspectives in the Escherichia coli proteome investigation. *Proteomics* 1: 409-23.

11. Hoogland, C., Sanchez, J.C., Tonella, L., Binz, P.A., Bairoch, A., Hochstrasser, D.F., and Appel, R.D. (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res* **28**: 286-288.

12. Karlin, S., Mrazek, J., Campbell, A., and Kaiser, D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J Bacteriol* **183**: 5025-40.

13. Opiteck, G.J., Ramirez, S.M., Jorgenson, J.W., and Moseley, M.A., 3rd (1998) Comprehensive two-dimensional high-performance liquid chromatography for the isolation of overexpressed proteins and proteome mapping. *Anal Biochem* **258**: 349-61.